Dan Janies

Biomedical Informatics
Ohio State

Comparative Genomics of Pathogens

janies-1@medctr.osu.edu

Examples:

Phylogeny of SARS associated Coronaviruses

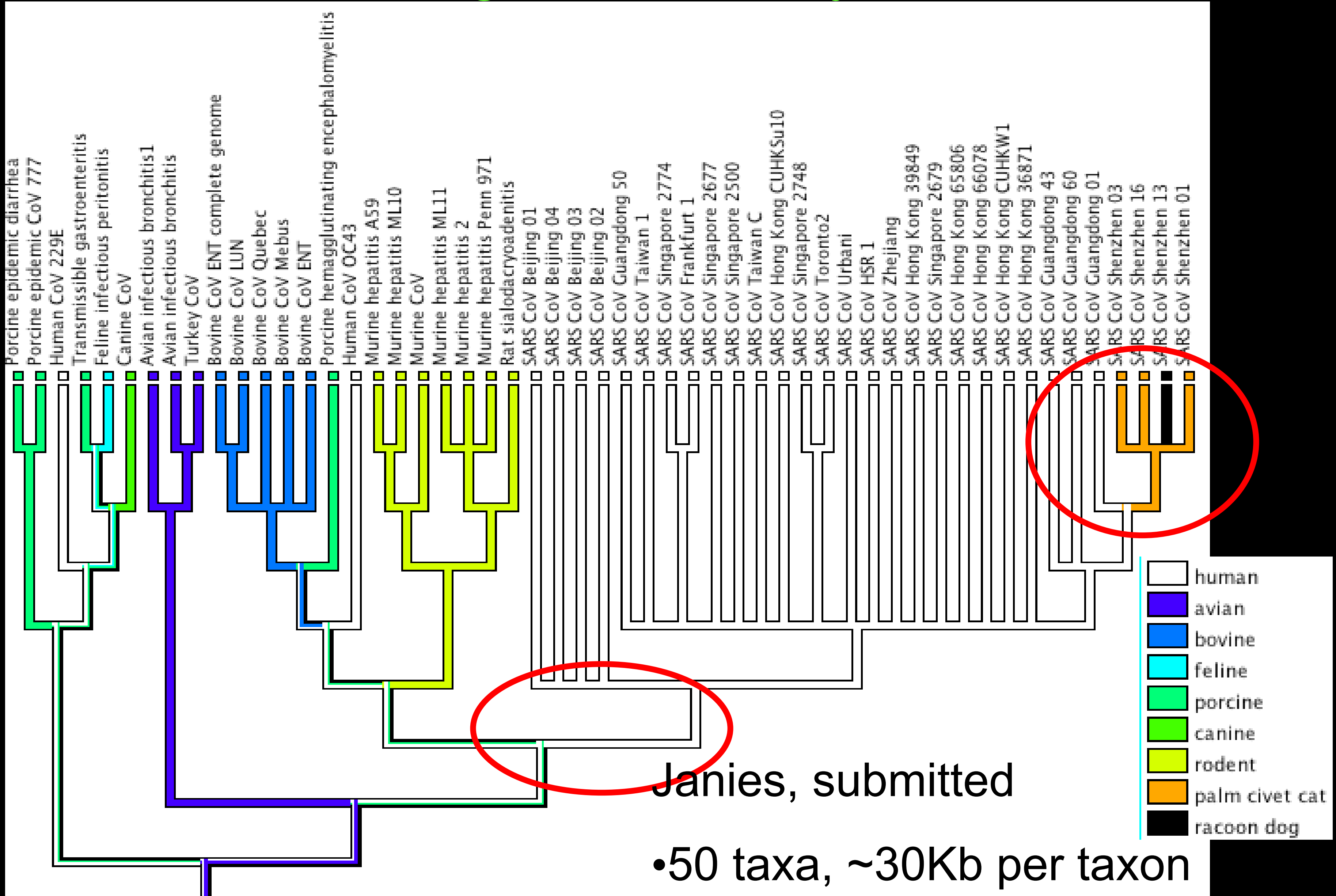
Software:

Intersection-exclusion analysis on trees

Hardware-software synergy:

Very fast multiple alignment and tree search in self-built parallel computing clusters

Host Switching Among Coronaviruses e.g. SARS Based on whole genome analysis



Janies, submitted

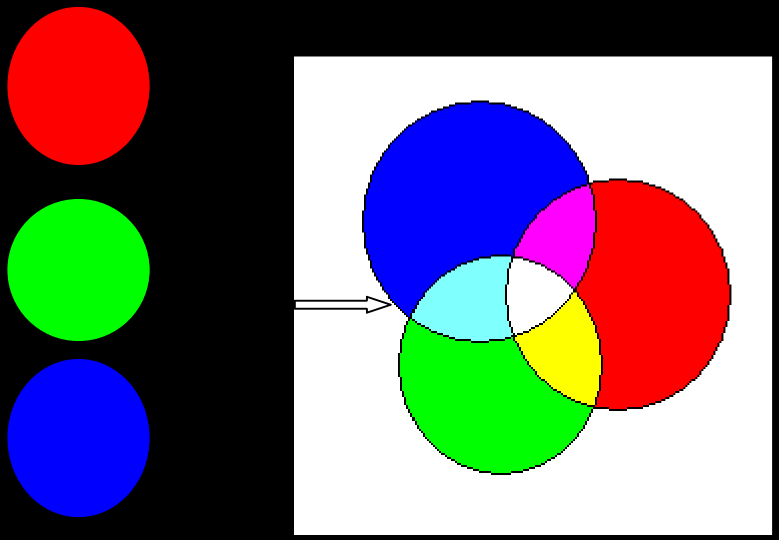
- 50 taxa, ~30Kb per taxon

Intersection-Exclusion Analysis on database of character change implied by trees.

node1-node2

node1-node3

node2-node3



Branches

Isolated Character change lists

Intersections and exclusions calculated

- 🟥 EXCL-node1-node2
- 🟦 EXCL-node2-node3
- 🟩 EXCL-node1-node3
- 🟨⬜ INTXnode1-node2Xnode1-node3
- 🟪⬜ INTXnode1-node2Xnode2-node3
- 🟦⬜ INTXnode1-node3Xnode2-node3
- ⬜ INTXnode1-node2Xnode1-node3Xnode2-node3

Shared changes in origin of SARS CoV in humans and subsequent infection of small carnivores

position	locus	function	anc	des	change
1893	nsp2-pp1a/pp1ab	two-proteases	A	C	Tv
1893	nsp2-pp1a/pp1ab	involved in transcriptional regulat	G	T	Tv
3310	nsp3-pp1a/pp1ab	coronavirus-host interactions	G	C	Tv
3310	nsp3-pp1a/pp1ab		T	C	Ti
6440	nsp3-pp1a/pp1ab		A	G	Ti
6440	nsp3-pp1a/pp1ab		G	T	Tv
22172	glycosylation site of Spike Prote	recognition of host cell receptor	C	A	Tv
22172	glycosylation site of Spike Protein		K	-	Del
22951	Spike Protein		C	G	Tv
22951	Spike Protein		Y	-	Del
23310	Spike Protein		B	C	ABC
23310	Spike Protein		T	C	Ti
25508	hypothetical protein sars3a		K	-	Del
25508	hypothetical protein sars3a		T	A	Tv
25544	hypothetical protein sars3a		C	T	Ti
25544	hypothetical protein sars3a		R	A	ABC
25844	hypothetical protein sars3a		B	G	ABC
25844	hypothetical protein sars3a		W	A	ABC

Tree search is a NP-complete problem

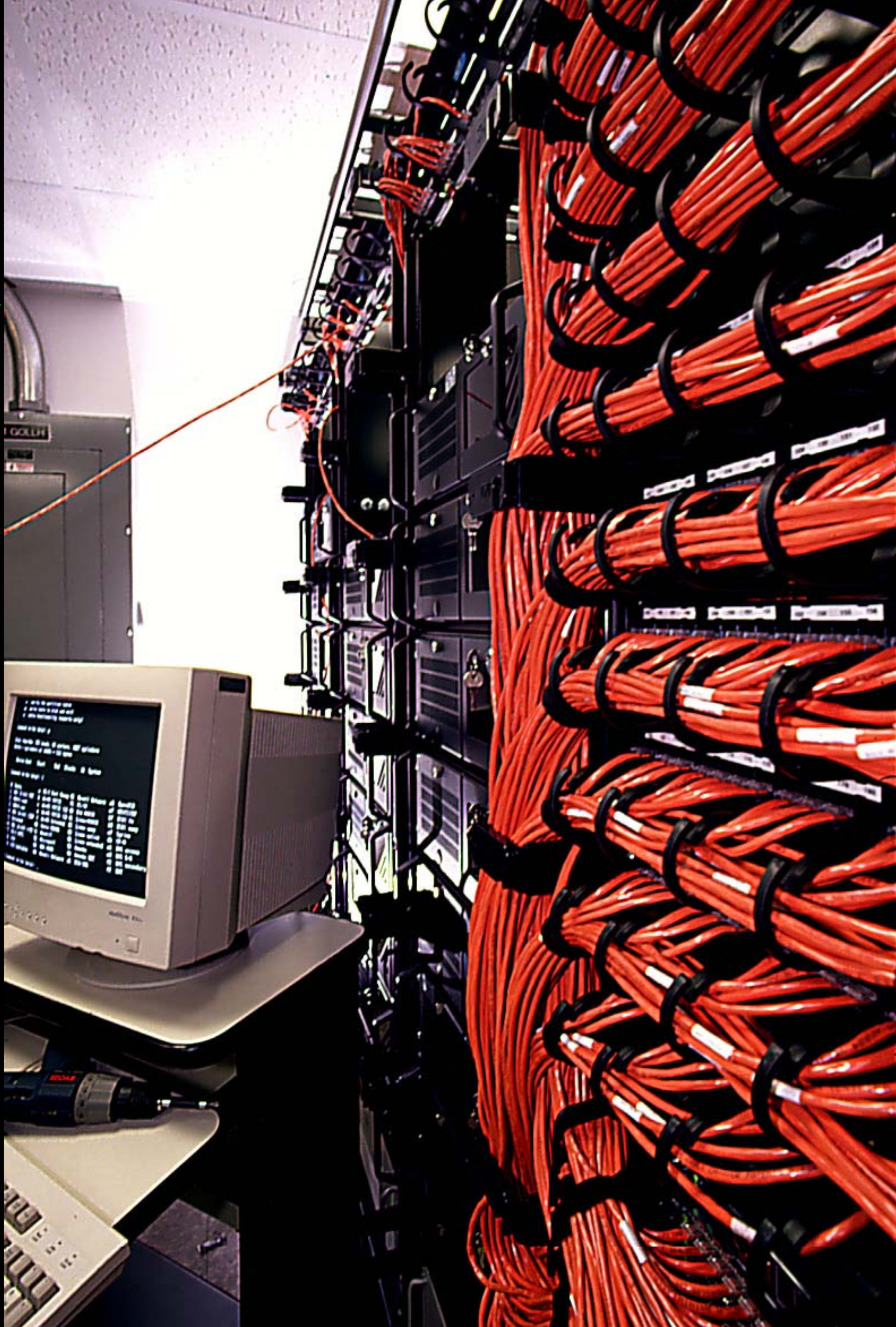
Group

$$y=\prod_{i=3}^{t}(2t-3)$$

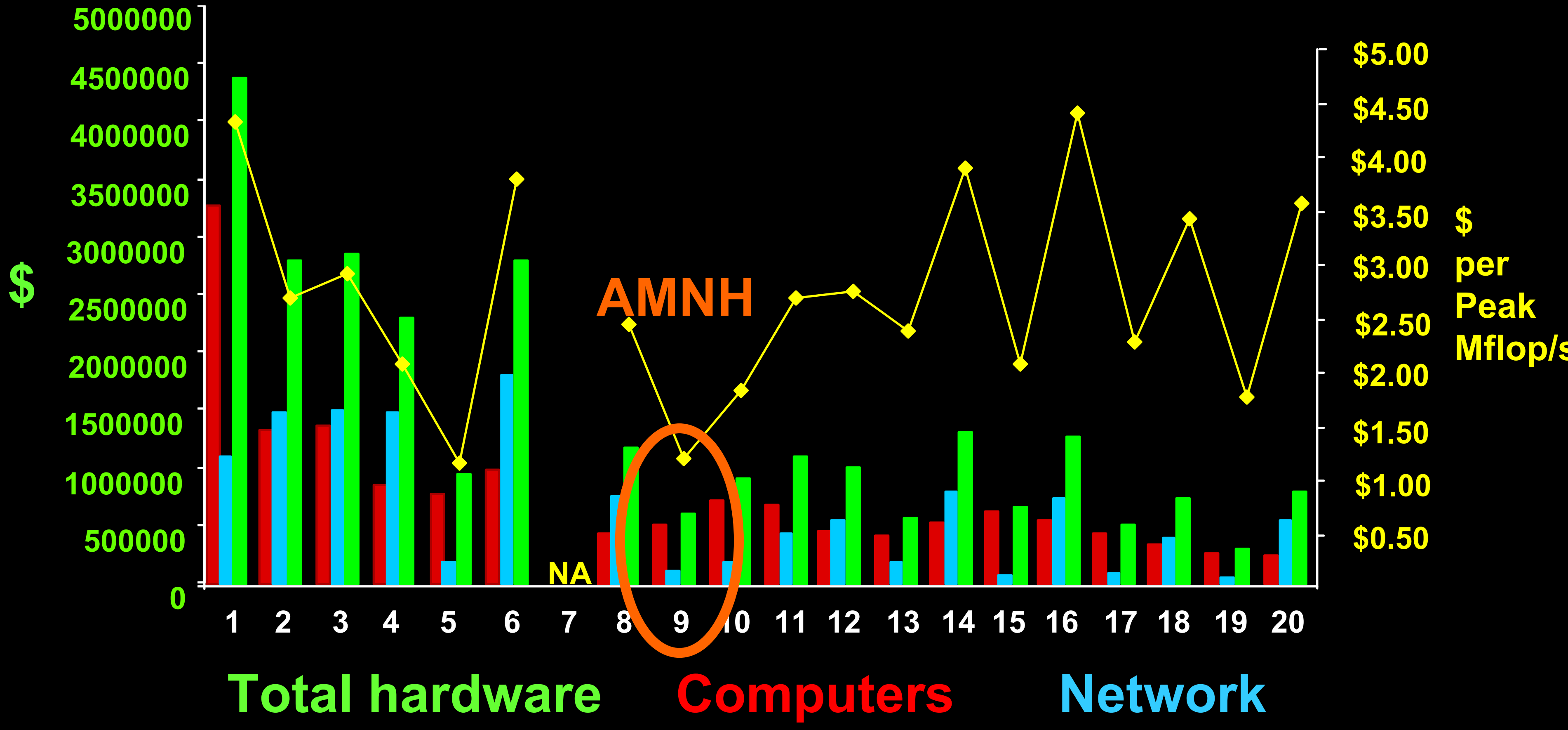
Heuristic Search Strategies

Monte Carlo	random tree building
Hill climbing	branch swapping
Simulated annealing	ratcheting
Genetic algorithms	tree fusion

564 processor
self-built cluster.
9th fastest cluster
in the world
6/2001

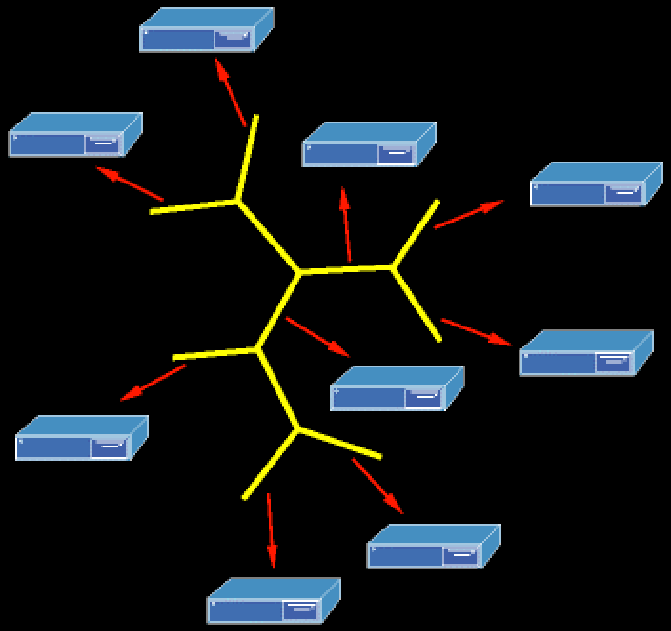


Networks explode the cost



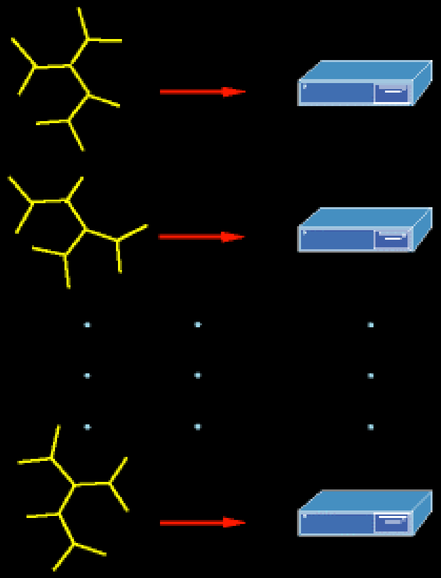
Ranking in June 2001by clusters.top500.org

Parallel building



High ratio of communication to computation

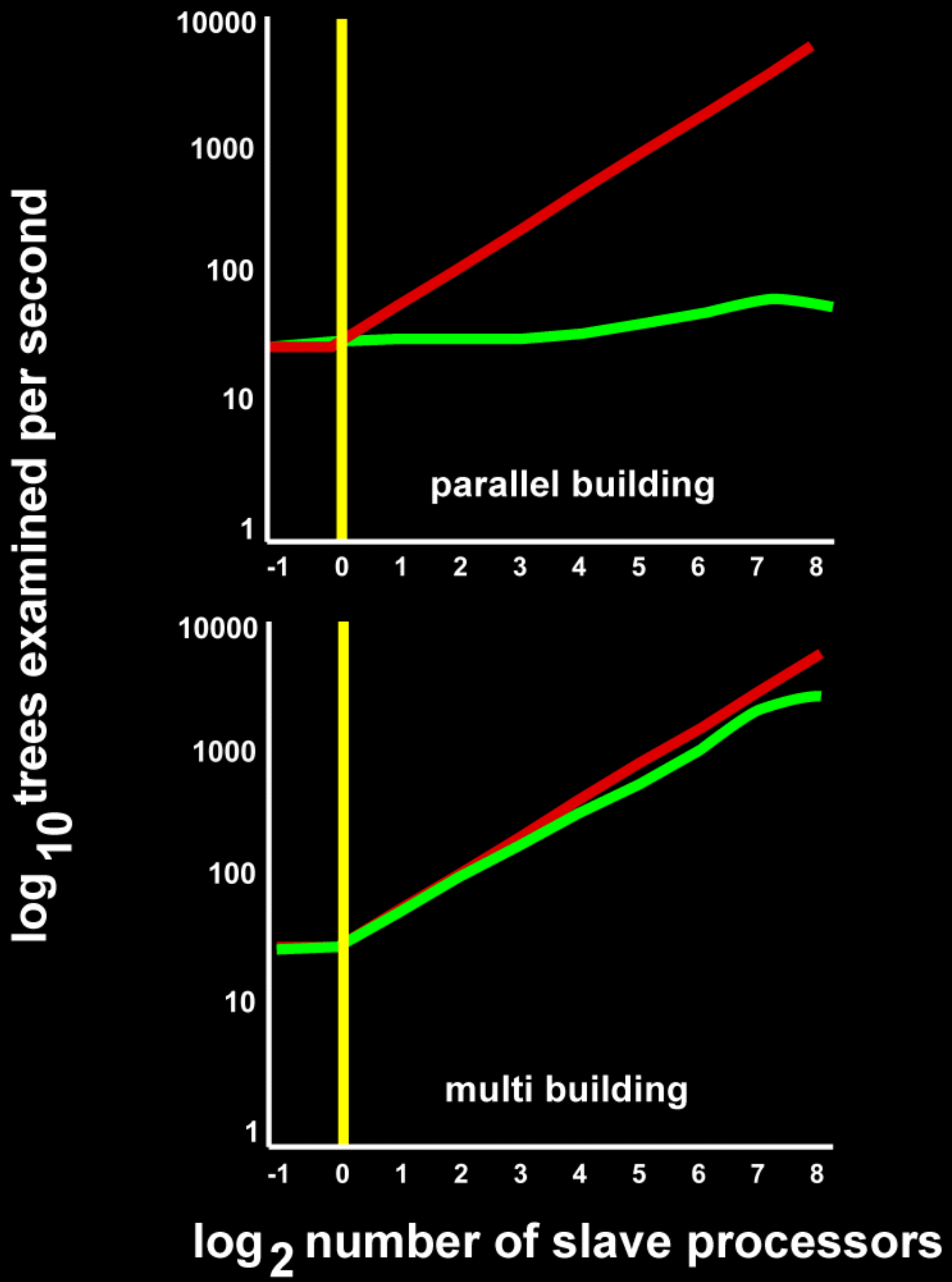
Multi building



Low ratio of communication to computation

Tree-based alignment:

speedup and granularity



More than just trees:

-Database of ancestor descendant changes

-Tools to search for Independently evolved genomic changes among diverse pathogens to provide well corroborated arguments for regions that confer pathogenicity or transmissibility

-Scalable and economic approaches to large datasets

-Sequencing coronaviruses before, during, and after host shifts